Baird, J. A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy and Practice*, *24*(1), 44-59. https://doi.org/10.1080/0969594X.2015.1108283

Rater Accuracy and Training Group Effects in Expert and Supervisor-Based Monitoring Systems

Jo-Anne Baird,[a*] Michelle Meadows,[b] George Leckie[c] and Daniel Caro[a]

[a]Oxford University Centre for Educational Assessment, Department of Education, University of Oxford, Oxford, UK. OX2 6PY.

[b]Ofqual, Spring Place, Herald Avenue, Coventry, UK. CV5 6UB.

[c]Centre for Multilevel Modelling, Graduate School of Education, 2 Priory Road, Bristol, UK. BS8 1TX.

* Corresponding author

jo-anne.baird@education.ox.ac.uk

Acknowledgements

The authors would like to thank the following people for comments on earlier drafts of this article: Dr Tom Bramley (Cambridge Assessment), Professor Paul Newton (Ofqual), Dr Edward Wolfe (Pearson) and the two anonymous reviewers.

Rater Accuracy and Training Group Effects in Expert and Supervisor-Based Monitoring Systems

ABSTRACT

This study evaluated rater accuracy with rater-monitoring data from high stakes examinations in England. Rater accuracy was estimated with cross-classified multilevel modeling. The data included face-to-face training and monitoring of 567 raters in 110 teams, across 22 examinations, giving a total of 5,500 data points. Two rater-monitoring systems (Expert consensus scores and Supervisor judgment of correct scores) were utilized for all raters. Results showed significant group training (table leader) effects upon rater accuracy and these were greater in the expert consensus score monitoring system. When supervisor judgment methods of monitoring were used, differences between training teams (table leader effects) were under-estimated. Supervisor-based judgments of raters' accuracies were more widely dispersed than in the Expert consensus monitoring system. Supervisors not only influenced their teams' scoring accuracies, they over-estimated differences between raters' accuracies, compared with the Expert consensus system. Systems using supervisor judgments of correct scores and face-to-face rater training are, therefore, likely to under-estimate table leader effects and over-estimate rater effects.

Keywords: rater accuracy, multilevel modelling, rater monitoring, rater training, table effects

Rater Accuracy and Training Group Effects in Expert and Supervisor-Based Monitoring Systems

INTRODUCTION

Accuracy in rating of students' performances for high-stakes examinations is typically monitored, so that assessment organisation staff can take decisions about raters' scoring performances. When poor accuracy is identified, actions might be taken on individual students' scores, raters might be subjected to further training or the rater's employment might be terminated. Rater accuracy is the most studied of a group of rater effects, which also include leniency or severity, 'halo' effects, central tendency and consistency. By rater accuracy, we mean the absolute difference between the rater's score and the correct score. In this study, we investigated accuracy effects. Little research has been conducted on the monitoring systems themselves that are used to evaluate raters and rating quality. This paper presents such a study and relates it to the effects of group training upon raters in face-to-face environments.

Two rater quality check systems used simultaneously in high stakes examinations in England were compared (Table 1). Both systems were applied across all of the examinations and raters. In the 'Expert-based monitoring system', correct scores were generated by the Principal Examiner and in the 'Supervisor-based monitoring' system, correct scores were generated by the rating team supervisors. The study involved a within-group comparison, as the raters and supervisors used both monitoring systems.

Individual students' work included in the monitoring was rated under only one system. One of our research questions was whether the different quality assurance systems would produce the same rank-ordering of raters' accuracy, as researchers have recently been interested in the stability of measures of rater effects over time or over subjects (e.g.,

Congdon & McQueen, 2000; Hoskens & Wilson, 2001; Harik et al., 2009; Lamprianou, 2006; Myford & Wolfe, 2009). The data included 22 examinations, with different styles of scoring rubric. 'Analytic' rubrics require scores for different elements of the response, which are aggregated, whilst 'holistic' rubrics require a single judgment of an overall score, even if that single judgment takes into account performances on elements of the response. In keeping with previous literature (e.g., Hartog and Rhodes, 1936; Klein et al., 1998; Chi, 2001), we also expected a significant mean difference between analytic and holistic rubrics, with greater accuracy for analytic scoring. Correct scores are easier to define with analytic rubrics in subjects with less ambiguous answers, such as mathematics compared with subjects such as English with more debatable 'correct' scores and holistic rubrics.

<Insert Table 1 >

Team training effects

There are few studies on the implications of training raters in teams, although this is common practice. In a study of quality checks on a mathematics high school examination, Wilson & Case (2000) found significant supervisor group training ('table leader') effects, but could not establish an association with the subsequent behaviour of raters within teams. Hoskens & Wilson (2001) also found supervisor effects in their analysis of rating quality checks for a high school Economics examination, with one team of raters being significantly more severe. They acknowledged that multilevel modelling (Raudenbush and Bryk, 2002; Goldstein, 2011; Snijders and Bosker, 2012) would be a better approach for estimating team training effects in rater accuracy with data of this kind (p. 134) and this is the analysis method used in the current research. Specifically, we estimated cross-classified multilevel models (Leckie,

2013; Rasbash and Goldstein, 1994; Raudenbush, 1993), which explicitly recognised the non-hierarchical aspects of the data.

Aims

Previous studies have shown that concealing prime scoring reduces the agreement level between the prime and second raters (McVey, 1975; Murphy, 1979). Here, we investigated the following research questions:

- Would there be stable rater accuracy effects across the two monitoring systems?
- Does analytic scoring produce greater rater accuracy?
- Will team training accuracy effects be observed?
- Will the Supervisor-based monitoring system produce greater observed rater accuracy?

METHOD

Assessments

A-level examinations are normally taken at age 18 (year 13) of high school following a two-year subject-based course. They are the main qualifications taken for access to university in England. First available in 1951, assessment processes are well established. Naturally, there have been many reforms to the curriculum, assessment design and administrative processes in their history. Assessment formats vary, but those included in the study were all written examinations, taken under controlled conditions. Questions were either short answer or extended answer and rubrics varied from generic descriptors to model answers. Essay-style questions were common.

Raters and training

A-level raters are usually practicing teachers of the subject and of the curriculum being assessed. Most raters have successfully scored an A-level assessment previously. Supervisors are selected on the basis of their ability to score to an acceptable standard (as measured through monitoring systems), as well as their aptitude for managing a team and typically have a number of years' scoring experience for a specific examination. The Principal Examiner is normally selected following national advertisement of the post, typically has a number of years' experience in scoring or examining, and is often a senior member of staff in a school.

In this study, face to face training took place on a single day for each question paper. Raters were organised into teams of approximately five and a supervisor headed each team. Principal Examiners gave introductory presentations to train the raters on interpretation of the question papers and rubrics and this was followed by team discussion of scoring students' work, using exemplar student performances which contained the Principal Examiners' correct scores that had also been agreed with the supervisors in the prior supervisor training event.

Data

Data were collected from Assessment and Qualifications Alliance A-level 2003 examination scoring. Scoring was paper-based and monitoring was conducted on paper, with scripts being sent to supervisors through the post. Data from the training phase of scoring were collated and analysed: raters were required to submit ten candidates' work (scripts) to their supervisors, who re-scored them, giving feedback to the raters and taking a decision about the quality of their prime scoring. Although ten scripts were normally scored, the sample was extended where there were doubts about a rater's performance. All data from examinations using both of these systems with large entries were included.

Five of the ten scripts for each rater were from the Expert-based system: photocopied scripts, with correct scores produced by the Principal Examiner. These photocopied scripts were also used to train the supervisors and were intended as a method of ensuring consistency across the population of supervisors. Obviously, the correct scores were not revealed during the raters' training, but the principles of scoring the work were discussed. The remaining sample of five scripts for each rater was from the Supervisor-based system: original scripts, double scored by the rater's supervisor. If extended sampling of a rater's scoring needed to be carried out, additional original scripts were requested from the raters. If doubts remained about the quality of raters' scoring, raters' employment might be terminated at this stage. All of the data, from successful and unsuccessful raters were included in the study. Further operational checks on the scoring were conducted at later stages in the process, but did not form part of this research.

A total of 5,500 quality assurance checks were collected for the study across 22 question papers, 110 supervisors and 567 raters. To facilitate comparison across question papers, absolute score differences between the scores allocated to the scripts by the raters and the supervisors were calculated and then converted to percentages of the maximum score for that paper. For example, a discrepancy of one score point resulted in a five percentage point difference in English Literature (where there was a maximum score of 20) but only a one percentage point difference in Physics (where there was a maximum score of 100). Absolute score differences were used to represent accuracy; score differences would have indicated severity and leniency, which was not the focus of the study.

Samples of students' work in the Expert-based monitoring system

The Principal Examiner for each question paper selected students' work for the photocopied scripts. The sample of these was selected to generate good training material: covering a range

of scores and optional questions where appropriate. However, the photocopied scripts were selected under time pressure and there were just five of them for each question paper. Although a non-random process, the Principal Examiners attempted to select representative examples of students' work.

Samples of students' work in the Supervisor-based monitoring system

Original scripts for re-scoring were selected by raters. Raters were likewise asked to submit scripts that represented a range of scores. It is possible that raters selected their sample of five original scripts carefully, to portray their scoring performance favourably. However, this was not a practical proposition, as raters were also under time pressure - to have their scoring vetted and be cleared to continue scoring. Further, Pinot de Moira et al. (2002) found very little change in rater scoring accuracy from the first to the final samples of A-level raters' scoring and Leckie and Baird (2011) reported similar findings for national examinations in England. So, original scripts included in the study were expected to be scored equally well as those scored later in the rater's allocation of students' work, despite the fact that sampling of scripts scored later in the marking period was not always conducted by the rater.

Analysis

Figure 1 displays the complex multilevel data structure using a classification diagram (Browne et al., 2001) separately for the Expert-based (Figure 1a) and Supervisor-based (Figure 1b) quality control systems. In the figure, the absolute score differences form level 1 of the data hierarchy. For each quality control system, there are five absolute score differences per rater and so absolute score differences are said to be nested within raters at level 2 in the data hierarchy. However, a complication arises in the Expert-based quality

control system as, for all raters on any given paper, the five absolute score differences related to the same five photocopied scripts. Thus, these absolute score differences are also separately nested within photocopied scripts at level 2 in the data hierarchy. This is not the case for original scripts, as each original script was scored by a different rater. Photocopied scripts are said to be cross-classified with raters at level 2 while the original scripts are confounded with the absolute score differences at level 1. In both quality control systems, raters are then nested within teams at level 3 where each team is led by a supervisor. Finally, teams are nested within question papers at the highest level in the data hierarchy, level 4.

<Insert Figure 1.>

A cross-classified multilevel model was fitted to the data (Leckie, 2013; Rasbash and Goldstein, 1994; Raudenbush, 1993). The response variable was the absolute score difference, which had a positive skew. Therefore, it was modelled on the log scale (Tabachnick and Fidell, 2007) so that the multilevel random effects better approximated the normality assumptions of the model (Goldstein, 2011). The log of zero is undefined and so a small positive constant of 0.5 was added to the response to facilitate the analysis. Choosing different values of this constant can be expected to yield slightly different model results. This, however, is readily studied, and in our analysis altering this constant makes little difference to any substantive conclusions. A description of the fixed and random parts of the model is presented in the Appendix.

Model fixed effects regression coefficients allowed us to test the hypotheses that score discrepancies were smaller for the Supervisor-based system than for the Expert-based system and that they were larger for subjects with holistic rubrics than those with analytic rubrics. The random part of the model included separate random effects variance components

for the two systems. With that, it was possible to examine the extent to which the degree of scoring accuracy varied across training teams and the extent to which it varied across the raters within their teams. Further, the variance components allowed us to examine whether there were differences in these patterns across the two rater monitoring systems. We fitted all models in the MLwiN multilevel modelling software (Rasbash et al. 2009) where we called MLwiN from within Stata (StataCorp, 2015) using the runmlwin command (Leckie and Charlton, 2013).

RESULTS

Descriptive analysis

Table 2 presents the mean absolute score difference for each question paper separately for the two systems. The table shows that, for 17 out of the 22 question papers, the mean absolute score differences were, on average, lower for the Supervisor-based system than for the Expert-based system. Thus, raters, on average, scored closer to the correct scores when they were set by supervisors (i.e., original scripts) than when the correct scores were set by the Principal Examiner in the Expert-based system (i.e., photocopied scripts). Supervisors did not double score blind and so the correct score that they assigned to each original script was very likely influenced by the rater's prime score, as seen in previous studies (McVey, 1975; Murphy, 1979).

<Insert Table 2>

Table 2 additionally shows that mean absolute score differences were notably higher for subjects with holistic rubrics than for subjects with analytic rubrics. Across both monitoring systems, the mean absolute score difference for subjects with analytic rubrics

ranged from just 1.1 to 3.4 while for subjects with holistic rubrics they ranged from 2.4 to 9.8. Thus, raters, on average, scored further from the correct scores on question papers with holistic rubrics than on question papers with analytic rubrics.

Multilevel analysis

Table 3 reports model results. In the fixed part of the model, taking the exponential of the intercept (0.663) (and subtracting 0.5) gives the median absolute score difference for the Expert-based system with an analytic rubric. This predicted value is 1.4 (exp(0.663)-0.5) and so the median absolute score difference was 1.4 percentage score points away from the correct score set by the Principal Examiner. The median absolute score difference for the three other combinations of script and rubric type can be similarly predicted and are 1.0 (exp(0.663-0.229)-0.5) for the Supervisor-based system with analytic rubrics, 2.7 (exp(0.663+0.490)-0.5) for the Expert-based system with holistic rubrics, and 2.0 (exp(0.663-0.229+0.490)-0.5) for Supervisor-based system with holistic rubrics. However, the interaction between monitoring system and rubric type is not presented in the model, as it was non-significant.

<Insert Table 3.>

The main effect for the Supervisor-based system (-0.229) is negative and significant (χ^2_1=6.71, p=0.010), while the main effect for holistic rubrics (0.490) is positive and significant (χ^2_1=14.08, p=0.001). These results, which take into account the complex clustering in these data, support those indicated by the descriptive statistics presented in Table 2: absolute score differences are significantly smaller for the Supervisor-based system compared with the Expert-based system and they are significantly larger for scripts with

holistic rubrics compared to those with analytic rubrics. Furthermore, these findings are as anticipated.

The random part of the model decomposes the variation in log absolute score differences that is unexplained by the fixed effects into separate variance components due to papers, training teams, raters, scripts and residual error variation. This decomposition is done separately for each monitoring system. The extent to which a given variance component is large compared to the other components of variation indicates the relative contribution of that component to the overall unexplained variation. To ease this interpretation of the variance components, we therefore present them in Table 4 together with variance partition coefficients (each component of variation divided by the total variation).

<Insert Table 4.>

We start by interpreting the variance components for the Expert-based system. We will then interpret the variance components for the Supervisor-based system and contrast these results with each other. For the Expert-based system, Table 4 shows that the paper variance (0.106) accounts for 7% of the unexplained variation in log absolute score differences. Thus, some question papers were more accurately scored than others even after adjusting for the large fixed effect of rubric type.

The training team variance (0.224) is large and accounts for 16% of the unexplained variation in log absolute score differences; some teams scored substantially more accurately than others. Importantly, this variability in team accuracy was not due to differences in supervisors' re-scoring standards as supervisors did not set the correct scores in the Expert-based system. This suggests that during training, supervisors created scoring cultures within their teams that differed in scoring standards from that specified by the Principal Examiner.

The rater level variance (0.082) accounts for just 6% of the unexplained variation in log absolute score differences. Thus, raters' average log absolute score discrepancies did not

vary much from other raters in their teams. Furthermore, the rater variance of 0.082 was considerably lower than the training team variance of 0.224. This indicates that the degree of scoring inaccuracy varied much more across teams than it did across the raters within these teams. In other words, raters scored relatively similarly within teams.

The script level variance (0.026) is very small and accounts for only 2% of unexplained variation in log absolute score differences. Thus, having adjusted for question paper, training team and rater effects, scripts differed very little in how accurately they were scored. That is, average score discrepancies were fairly similar across scripts and so all scripts were scored with broadly the same degree of accuracy.

The residual error variance (0.998) accounts for 69% of the unexplained variation and, as is common in studies of scoring accuracy, is by far the largest component of unexplained variation (see, for example, Leckie and Baird, 2011 and Pinot de Moira et al., 2002). Idiosyncrasies of the interaction between raters and students' performances are represented by the residual error variance.

Table 4 shows that the estimated variance components for the Supervisor-based system differed substantially from those for the Expert-based system. However, only the training team variances and rater variances differ significantly by system (χ^2_1=7.04, p=0.008 and χ^2_1=5.77, p=0.016 respectively); the differences observed for the question paper variances and the residual variances are not statistically significant (χ^2_1=1.30, p=0.255 and χ^2_1=1.58, p=0.209, respectively). We therefore limit our discussion to the training team and rater variance components.

For the Supervisor-based system, the training team variance (0.103) is significantly smaller than that for the Expert-based system (0.224). In contrast, the rater variance (0.157) for the Supervisor-based system is significantly larger than that for the Expert-based system (0.082). Thus, in the Supervisor-based system, when supervisors set the correct scores, there was

significantly less variation in absolute score differences between teams, but significantly more variation between raters, within their teams. In other words, when supervisors set the correct scores, teams appear more similar to one another than when the Principal Examiners set the correct scores. However, supervisors judged their raters to vary considerably more in scoring accuracy when they set the correct scores than when the Principal Examiner set the correct scores.

Table 3 also presents the covariance between the training team effects for the two systems. The covariance (0.063) is positive and statistically significant (χ^2_1=6.63, p=0.010), however the corresponding correlation of 0.41 (0.063/√(0.244x0.103)) is moderate. Thus, teams which scored more accurately than average in the Expert-based system more often than not scored more accurately than average for the Supervisor-based system and vice versa, but this would by no means always be the case. Table 3 also presents a positive and statistically significant covariance (0.068) between the rater effects for the two systems. The corresponding correlation of 0.60 (0.068/√(0.082x0.157)) is moderately high and so raters that, within their teams, scored more accurately than average under one monitoring system tended to also do so under the other system. However, the fact that the correlation is not higher again highlights that the choice of quality control mechanism could lead to inconsistent descriptions of raters' levels of scoring accuracy.

Figure 2 presents predicted absolute score differences for teams, together with 95% confidence intervals, separately for the two systems. We centre these predictions around the median predicted absolute score difference in each system. Note that these predictions reflect differences between teams which remain, even after adjusting for question paper, rubric type and all other factors in the model. The figure clearly shows that, not only is the predicted median absolute score difference for the Supervisor-based system less than it is for the Expert-based system (denoted by the horizontal lines), but the predicted absolute score

differences for teams vary less for the Supervisor-based system than for the Expert-based system. The figure also shows that the vast majority of team predicted absolute score differences were not significantly different from the median team (denoted by the horizontal lines), although more were significantly different from the median for the Expert-based system. The imprecision of these team predictions largely reflects the small size of teams: the average team had only five raters.

<Insert Figure 2.>

Figure 3 presents predicted absolute score differences for the raters, together with 95% confidence intervals. Again we centre these predictions around the median predicted absolute score difference in each system. These predictions reflect differences between raters even after we have accounted for their team membership and all other factors in the model. The figure clearly shows that the predicted absolute score differences for raters vary more for the Supervisor-based system (original scripts) than for the Expert-based system (photocopied scripts). However, almost none of the rater predicted absolute score differences were significantly different from the median rater (denoted by the horizontal lines). This reflects the small number of scripts scored by each rater (approximately five).

DISCUSSION

We take each of our research questions in turn. First, only moderate stability of rater effects ($r=0.6$) was found across the two monitoring systems, somewhat worryingly suggesting that different impressions of rater performance could be given by the adoption of a particular system. Other studies too have shown instability in rater effects (Baird et al., 2013; Congdon & McQueen, 2000; Hoskens & Wilson, 2001; Harik et al., 2009; Lamprianou, 2006; Myford & Wolfe, 2009), which might be explained by small sample sizes in the monitoring checks. This study is the first to show instability across monitoring systems.

Second, as expected, analytic scoring generally produced greater scoring accuracy. Third, significant training group effects were found under both monitoring systems. This is the first study to show this as a general effect, rather than for a particular team, and it is the first to use multilevel modelling to do so. Fourth, as anticipated, the Supervisor-based monitoring system appeared to produce greater observed accuracy, which we explain by the fact that in this system supervisors can be influenced by the original scores given to students' work.

Interestingly, and unexpectedly, there were larger group training effects in the Expert-based monitoring system and larger within-group spread of rater accuracy effects in the Supervisor-based monitoring system. Thus, it appears that when supervisors were permitted to decide the correct score in response to the original rater's score, they produced data that, compared with the Expert-based system, a) under-estimated overall rater inaccuracy, b) over-estimated differences in accuracy between raters and c) under-estimated differences between groups in terms of accuracy. The mechanisms of findings a) and c) are likely to be the biasing effect of the original rater's score, but b) is likely to be a bias caused by supervisor attempts to distinguish good and weak rater performances within their teams. However, the mechanism for production of group training effects under the Expert-based monitoring system is as yet unexplained. After all, in that system, supervisors simply applied the correct score supplied by the Principal Examiner. We therefore concluded that group effects were caused during discussion with supervisors ('table leaders') at the training meetings. Note that the training materials contained correct scores created by the Principal Examiner (Expert), so the effects of team discussion must have been powerful.

As this study involved secondary data analyses of operational datasets, there were several limitations to the available data. Information regarding supervisors' views of raters, and supervisors' and raters' backgrounds might have added to the study, as it would have been possible to investigate how these factors might have influenced the extent to which

supervisors took into account the raters' views. This could have led to a rational explanation for the biasing effect of viewing the raters' original scores, for example.

Raters knew that their performance was being monitored. We did not have data on random checks of scoring performance for which the raters were unaware of scrutiny. Our findings are therefore only generalizable to scoring if the training check data included in the current study were fair indicators of rating and checking behaviour. As indicated previously, there is some evidence to suggest that this is a reasonable assumption (Leckie and Baird, 2011 and Pinot de Moira et al., 2002).

Our data included a small number of raters whose employment was terminated due to unacceptable levels of scoring accuracy. Therefore, the findings are likely to indicate higher levels of variability than the final operational scores.

Although there was variability in assessment formats and subjects included in this study, all of the data were drawn from A-level examinations taken in England, scored by experienced raters who were qualified teachers in the subjects. Research is needed to investigate whether the same effects are found in different contexts. Further, the small-scale nature of samples within particular subjects precluded their separate study, but it might be that Supervisor-based checks are sufficient for subjects such as mathematics with unambiguous, analytic rubrics.

CONCLUSIONS

The combination of cohesion to the supervisor scoring standard across teams and spread of raters within teams is likely to be caused by supervisors discriminating between their team members. In effect, their monitoring indicated that they were more attuned with some of the team members' scoring than to others. Several processes could account for this, including a belief that their team represented a wider range of scoring abilities than was the case, or other

social and administrative factors influencing the degree to which the original marking had an influence upon the supervisors' monitoring. Availability of original scores in the Supervisor-based monitoring systems had a biasing effect that produced a lower estimate of rater deviance from correct scores, in the fixed effect. A conclusion from this study is that face-to-face training coupled with Supervisor-based monitoring is likely to under-estimate group training effects and over-estimate rater effects. These are important findings, as face-to-face training and Supervisor-based monitoring systems are still the norm in many examination settings for practical reasons. Although in many contexts monitoring of marking is still paper-based, two-thirds of marking in public examinations in England is now on-screen, in keeping with the expansion of marking technology in other settings, such as Hong Kong, Korea, Australia and the US (Ofqual, 2014). With the advent of technology in these systems, online training and Expert-based monitoring systems should produce more accurate depictions of rater severity effects.

REFERENCES

Baird, J.-A. , Hayes, M., Johnson, R., Johnson, S. & Lamprianou, I. (2013). *Marker effects and examination reliability. A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling* (Ofqual/13/5261). Coventry: Office of Qualifications and Examinations Regulation. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378059/2013-01-21-marker-effects-and-examination-reliability.pdf

Browne, W. J. (2009). *MCMC Estimation in MLwiN, v2.13*. University of Bristol: Centre for Multilevel Modelling.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling, 1*, 103–124.

Browne, W. J., Steele, F., Golalizadeh, M., & Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society. Series A, (Statistics in Society) 172,* 579–598.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement, 2,* 379–388.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37,* 163–178.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). New York, NY: Wiley.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46,* 43–58.

Hartog, P., & Rhodes, E. C. (1936). *The marks of examiners*. London: Macmillan.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38,* 121–145.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., … Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11,* 121–137.

Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. *Journal of Applied Measurement, 7,* 192–205.

Leckie, G. (2013). *Cross-Classified Multilevel Models - Concepts*. LEMMA VLE Module 12, 1-60. www.bristol.ac.uk/cmm/learning/module.../12-concepts-example.pdf.

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency and rater experience. *Journal of Educational Measurement, 48,* 399–418.

Leckie, G., & Charlton, C. (2013). runmlwin - A program to run the MLwiN multilevel modelling software from within Stata. *Journal of Statistical Software, 52*(11), 1–40.

McVey, P. J. (1975). The errors in scoring examination scripts in electronic engineering. *International Journal of Electronic Engineering Education, 12,* 203–216.

Murphy, R. J. L. (1979). Removing the scores from examination scripts before re-scoring them: Does it make any difference? *British Journal of Educational Psychology, 49,* 73–78.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46,* 371–389.

Ofqual (2014). *Review of quality of marking in exams in A levels, GCSEs and other academic qualifications.* Final Report (Ofqual/14/5379). Retrieved from:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf.

Pinot de Moira, A., Massey, C., Baird, J.-A., & Morrissy, M. (2002). Marking consistency over time. *Research in Education, 67,* 79–87.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). MLwiN (Version 2.1). University of Bristol: Centre for Multilevel Modelling. Retrieved from: http://www.mlwin.com

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19,* 337–350.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics, 18,* 321–349.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

StataCorp (2015). Stata Statistical Software (Release 14). College Station, TX: StataCorp LP.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in reader drift. In M. Wilson & G. Engelhard, Jnr (Eds.), *Objective measurement: Theory into practice* (Vol. 5, 113–134). Stamford, CT: Ablex Publishing.

APPENDIX

A cross-classified multilevel model was fitted to the data (Leckie, 2013). This model is written below using the 'classification' notation of Browne et al. (2001), which avoids a proliferation of subscripts when many random classifications are present.

$$y_i = \beta_0 + \beta_1 supervisor_i + \beta_2 holistic_i + \left(u^{(5)}_{0\text{question paper}(i)} + u^{(4)}_{0\text{training team}(i)} + u^{(3)}_{0\text{rater}(i)} + u^{(2)}_{0\text{script}(i)} + e_{0i}\right) expert_i + \left(u^{(5)}_{1\text{question paper}(i)} + u^{(4)}_{1\text{training team}(i)} + u^{(3)}_{1\text{rater}(i)} + e_{1i}\right) supervisor_i \quad \ldots(1)$$

$$\text{question paper}(i) \in \left(1, \ldots, \mathrm{J}^{(5)}\right), \qquad \text{training team}(i) \in \left(1, \ldots, \mathrm{J}^{(4)}\right)$$

$$\text{rater}(i) \in \left(1, \ldots, \mathrm{J}^{(3)}\right), \qquad \text{script}(i) \in \left(1, \ldots, \mathrm{J}^{(2)}\right)$$

$$i = 1, \ldots, \mathrm{N}$$

$$\begin{pmatrix} u^{(5)}_{0\text{question paper}(i)} \\ u^{(5)}_{1\text{question paper}(i)} \end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(5)} & \\ \sigma_{u01(5)} & \sigma^2_{u1(5)} \end{pmatrix}\right\}$$

$$\begin{pmatrix} u^{(4)}_{0\text{training team}(i)} \\ u^{(4)}_{1\text{training team}(i)} \end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(4)} & \\ \sigma_{u01(4)} & \sigma^2_{u1(4)} \end{pmatrix}\right\}$$

$$\begin{pmatrix} u^{(3)}_{0\text{rater}(i)} \\ u^{(3)}_{1\text{rater}(i)} \end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(3)} & \\ \sigma_{u01(3)} & \sigma^2_{u1(3)} \end{pmatrix}\right\}$$

$$u^{(2)}_{0\text{script}(i)} \sim \mathrm{N}\left(0, \sigma^2_{u0(2)}\right)$$

$$\begin{pmatrix} e_{0i} \\ e_{1i} \end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e0} & \\ 0 & \sigma^2_{e1} \end{pmatrix}\right\}$$

The response variable y_i is the i-th log absolute score difference. The model includes fixed effects for the intercept and for separate binary indicators for whether the script is from the

Supervisor-based system (supervisor = 1) or the Expert-based system (supervisor = 0) and whether the script has a holistic rubric (holistic = 1). The model has five classifications of random effects: absolute score difference, script, rater, training team and question paper. These classifications are numbered from one to five. Hence, the '(2)', '(3)', '(4)', and '(5)' superscripts and subscripts identify random effects that are associated with each of the classifications above the absolute score difference level where the '(1)' superscripts and subscripts are implicit for convenience of notation.

The classification function 'training team(i)' denotes the training team associated with the i-th absolute score difference. Training teams are indexed from 1 to $\mathrm{J}^{(4)}$ and $u^{(4)}_{0\text{training team}(i)}$ and $u^{(4)}_{1\text{training team}(i)}$ are the different effects that the training team has on the i-th absolute score difference for the Expert-based system and the Supervisor-based system. The classification functions and random effects for the other classifications are similarly defined.

All random effects are assumed (bivariate) normally distributed, independent across classifications and independent of any predictor variables included in the model.
For the Supervisor-based system, there is no script effect as each script is scored by only one rater. The script effect is therefore confounded with the level 1 residual. At level 1, the covariance is structurally zero as each absolute score difference belongs to either the Expert-based system or the Supervisor-based system, but not both.

Since 'classification' notation does not show the multilevel structure in the data, 'classification diagrams' are typically presented in addition to the model equation (Browne et al., 2001). Figure 1a depicts a classification diagram for the complex cross-classified model structure assumed for the Expert-based system in the model equation. Figure 1b depicts the corresponding classification diagram for the four-level hierarchy assumed for Supervisor-based system in the model equation. The diagrams have one node for each classification in

the model. Two nodes connected by an arrow indicate a nested relationship while two unconnected nodes indicate a crossed relationship.

Estimation and Software

The model was fitted using Markov chain Monte Carlo (MCMC) based algorithms as implemented in MLwiN (Browne, 2009; Rasbash et al., 2009). Initial values for all fixed part parameters and random part covariance parameters were set to zero, initial values for all random part variance parameters were set to 0.1. The model was run for a burn-in of 10,000 iterations followed by a monitoring period of 100,000 iterations. We used hierarchical centring (Browne, 2009; Browne et al., 2009) to produce chains that exhibit better mixing. We use the standard default prior distributions provided by MLwiN: diffuse uniform priors for the fixed part parameters and minimally informative inverse Wishart priors for the random part covariance matrices. Informal visual assessments of the parameter chains and standard MCMC convergence diagnostics suggested that the sampler was run for sufficiently long. The effective sample size for every parameter chain exceeded 250.

When we report the results, we present the means and standard deviations (SDs) of the monitoring iterations for each parameter. These quantities are analogous to the parameter estimates and standard errors obtained in frequentist analyses.

TABLES AND FIGURES

Table 1
Study 1 Rater Monitoring Systems

	Expert-based system	Supervisor-based system
Rater's original score	Observable	Observable
Correct score assignment	Principal Examiner	Supervisor
Selection of check sample	Principal Examiner	Rater
Check sample	Same 5 for all raters for an examination	Different 5 for each rater
Presentation of check sample	Photocopied student work from current examination	Original student work from current examination
Training	Face-to-face meetings on tables with the supervisor	
Scoring process	Entire student's script (all items)	

Table 2
Mean absolute score differences as a percentage of maximum mark

Rubric	Subject	Question Paper *n*	Expert-based System mean	Supervisor-based System mean	Teams *n*	Raters *n*	Checks *n*
Analytic	Biology	1	1.8	1.5	9	49	489
		2	2.1	1.6	8	40	398
		3	3.4	2.4	4	16	160
		4	2.9	1.8	7	29	289
	Environmental science	1	1.7	2.5	2	7	62
		2	2.7	1.6	2	7	70
	Maths	1	2.4	1.1	3	15	150
		2	1.4	1.6	5	24	240
		3	2.7	1.8	2	8	80
		4	2.2	1.7	3	14	139
	Physics	1	1.5	1.3	2	8	73
		2	2.5	2.6	2	11	110
		3	2.2	1.8	2	8	80
		4	2.2	1.6	2	9	90
Holistic	English literature	1	3.0	3.8	6	24	236
		2	9.7	6.2	14	78	754
		3	3.3	3.3	8	64	549
		4	4.9	4.5	9	40	384
	History	1	8.5	4.8	9	62	620
		2	5.3	2.4	3	8	80
		3	5.5	4.7	2	6	60
	Media Studies	1	9.8	5.8	6	40	387
Overall		22	4.3	3.2	110	567	5,500

Table 3
Cross-classified multilevel model for log absolute score differences

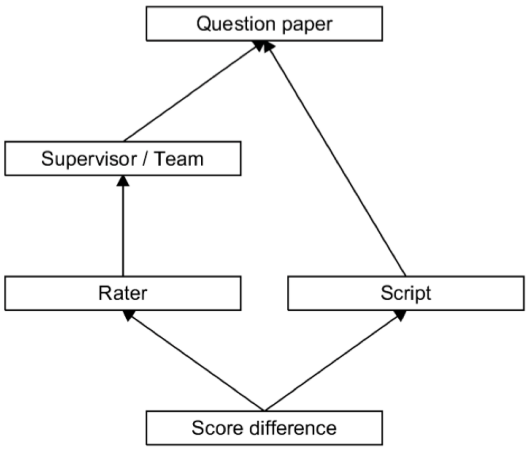
Parameter	Est.	SE	p
Fixed part			
Intercept	0.663	0.109	<0.001
Supervisor-based system (ref. Expert-based system)	-0.229	0.088	0.010
Holistic rubric (ref. analytic rubric)	0.490	0.131	<0.001
Random part			
Expert-based system: question paper variance	0.106	0.049	0.031
Expert- and Supervisor-based systems: question paper covariance	0.039	0.027	0.153
Supervisor-based system: question paper variance	0.053	0.025	0.034
Expert-based system: team variance	0.224	0.041	0.000
Expert- and Supervisor-based systems: team covariance	0.063	0.024	0.010
Supervisor-based system: team variance	0.103	0.026	<0.001
Expert-based system: rater variance	0.082	0.017	<0.001
Expert- and Supervisor-based systems: rater covariance	0.068	0.015	<0.001
Supervisor-based system: rater variance	0.157	0.026	<0.001
Expert-based system: script variance	0.026	0.011	0.013
Expert-based system: residual variance	0.988	0.030	<0.001
Supervisor-based: residual variance	1.043	0.031	<0.001

Note: p-values are based on standard Wald tests and are therefore approximate for random-part parameters.

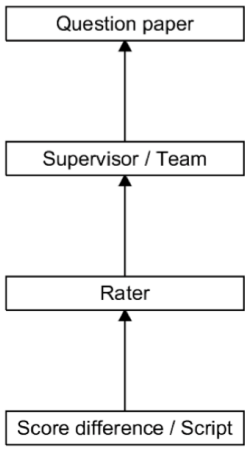
Table 4
Variance components and variance partition coefficients

Classification	Variance components		Variance partition coefficients	
	Expert-based system	Supervisor-based system	Expert-based system	Supervisor-based system
Question paper	0.106	0.053	0.07	0.04
Training team	0.224	0.103	0.16	0.08
Rater	0.082	0.157	0.06	0.12
Script	0.026	-	0.02	-
Residual	0.998	1.043	0.69	0.77
Total	1.436	1.338	1.00	1.00

Figure 1. Cross-classification data diagram: (a) Expert-based monitoring system (photocopied scripts) and (b) Supervisor-based monitoring system (original scripts).



(a)



(b)

Figure 2. Predicted absolute score differences for teams presented with 95% confidence intervals, plotted separately by system. The horizontal lines denote the median absolute score difference in each system.

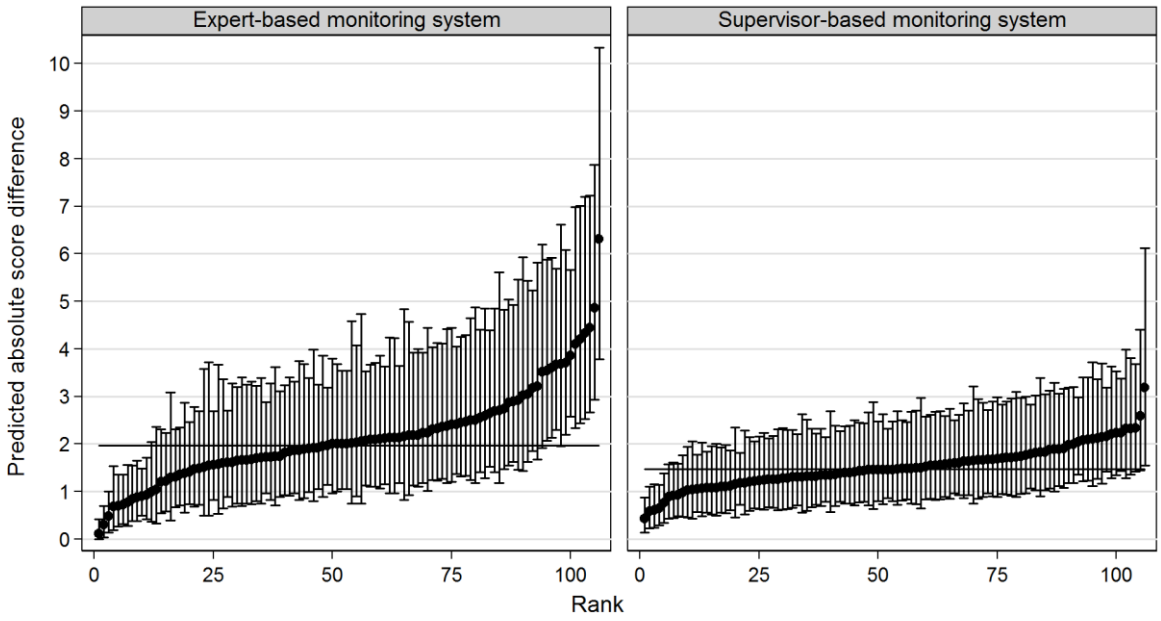


Figure 3. Predicted absolute score differences for raters presented with 95% confidence intervals, plotted separately by system. The horizontal lines denote the median absolute score difference in each system. Note that for graphical clarity we have only plotted every fifth rater.